Quantifying Uncertainty and Stability Among Highly Correlated Predictors: A Subspace Perspective

Xiaozhu Zhang

Department of Statistics
University of Washington

Joint with

Armeen Taeb
University of Washington

Jacob Bien
University of Southern California

Setup

Suppose that we have a dataset containing (X, Y), where

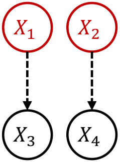
- $X \in \mathbb{R}^{n \times p}$ is a fixed design matrix for features
- Some features are **nearly linearly dependent**
- $Y \in \mathbb{R}^n$ is a vector for response variable

Model selection task

Which model $S \subseteq \{1, \ldots, p\}$ can explain the response variable?

Table of Contents

Toy example:

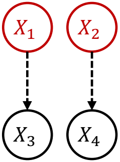


- $X_1 \approx X_3, \quad X_2 \approx X_4$
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$
- So the true support is: $S^\star = \{1, 2\}$

Three Challenges

Toy example:

- $X_1 \approx X_3, \quad X_2 \approx X_4$
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$
- So the true support is: $\quad S^\star = \{1, 2\}$



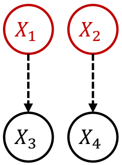
Challenge 1: Measuring Subspace Accuracy

How to define **True Positives (TP)** and **False Positives (FP)**?

Three Challenges

Toy example:

- $X_1 \approx X_3, \quad X_2 \approx X_4$
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$
- So the true support is: $\quad S^\star = \{1, 2\}$



Challenge 1: Measuring Subspace Accuracy

How to define **True Positives (TP)** and **False Positives (FP)**?

Suppose the selected set is: $\quad \widehat{S} = \{1, 4\}$

- Naively: 1 true positive, 1 false positive
- Since $X_4 \approx X_2$, is X_4 really "false"?

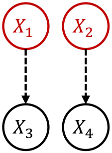
Goal: Redefine TP / FP so that

$$\mathrm{TP}(S^\star, \widehat{S}) \approx 2, \quad \mathrm{FP}(S^\star, \widehat{S}) \approx 0$$

Three challenges

Toy example

- X_1 and X_3 are highly correlated
  X_2 and X_4 are highly correlated
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$



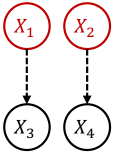
Challenge 2

How to quantify **stability** of selected features and sets?

Three challenges

Toy example

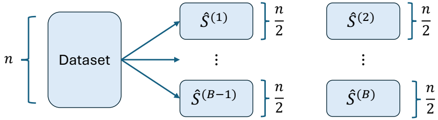
- X_1 and X_3 are highly correlated
  X_2 and X_4 are highly correlated
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$



Challenge 2

How to quantify **stability** of selected features and sets?

Stability selection (Meinshausen and Bühlmann, 2010):

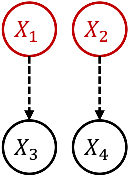


Stability of feature j:
- **selection proportion** $\pi(\{j\})$

Three Challenges

Toy example:

- X_1 and X_3 are highly correlated
  X_2 and X_4 are highly correlated
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$



Challenge 2

How to quantify **stability** of selected features and sets?

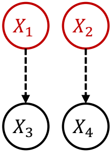
Stability selection (Meinshausen & Bühlmann, 2010):

- $\sim$50% of subsamples select X_1
- $\sim$50% of subsamples select X_3
- $\pi(\{1\}) \approx \pi(\{3\}) \approx 0.5 \quad \Rightarrow$ **Not stable!**

Three challenges

Toy example

- X_1 and X_3 are highly correlated
  X_2 and X_4 are highly correlated
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$



Challenge 2

How to quantify **stability** of selected features and sets?

Stability selection (Meinshausen and Bühlmann, 2010):

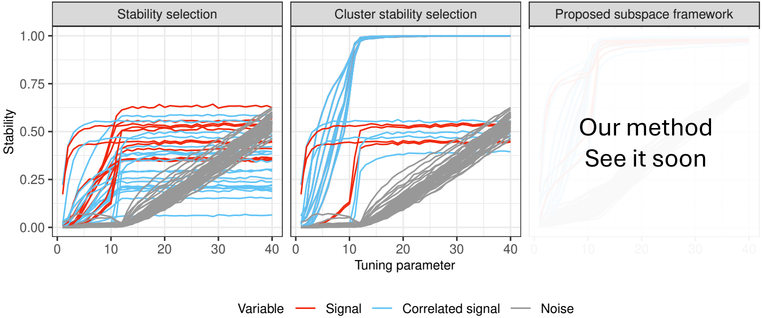
Use Lasso or ℓ_0 as the base procedure:

- $\sim$ 50% subsamples choose X_1
- $\sim$ 50% subsamples choose X_3
- $\pi(\{1\}) \approx \pi(\{3\}) \approx 0.5 \;\Rightarrow\;$ **NOT stable!**

Goal: re-define stability
s.t. $\pi(\{1\}) \approx \pi(\{3\}) \approx 1$

An experiment: Quantifying stability

A comprehensive synthetic dataset:

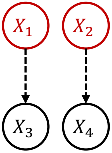


For stability selection (and its variant):
- Many signal and correlated signal features are **not stable**!

Three challenges

Toy example

- X_1 and X_3 are highly correlated
  X_2 and X_4 are highly correlated
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$



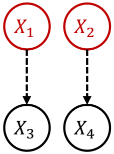
Challenge 3

How to **aggregate** features into a model?

Three challenges

Toy example

- X_1 and X_3 are highly correlated
  X_2 and X_4 are highly correlated
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$



Challenge 3

How to **aggregate** features into a model?

Suppose now all 4 variables are stable.

Can we have $\widehat{S} = \{1, 2, 3, 4\}$?

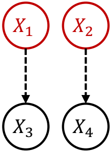
- No! Because X_1 and X_3 are redundant! (as well as X_2 and X_4)
- Instead, we want multiple models: $\{1, 2\}$, $\{1, 4\}$, $\{2, 3\}$, $\{3, 4\}$

Goal: an algorithm that outputs **multiple** models **without redundancy**

Beyond feature perturbation

Toy example ♠:

- X_1 and X_3 are highly correlated
  X_2 and X_4 are highly correlated
- $y = \beta_1^\star X_1 + \beta_2^\star X_2 + \epsilon$



We also need to deal with **more complicated** structures...

Toy example ♣:

- $X_3 = X_1 + X_2 + \delta$ (a perturbed **linear combination**)
- $y = X_1 - X_2 + \epsilon$

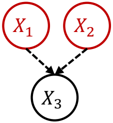


Table of Contents

Our contribution: a subspace perspective

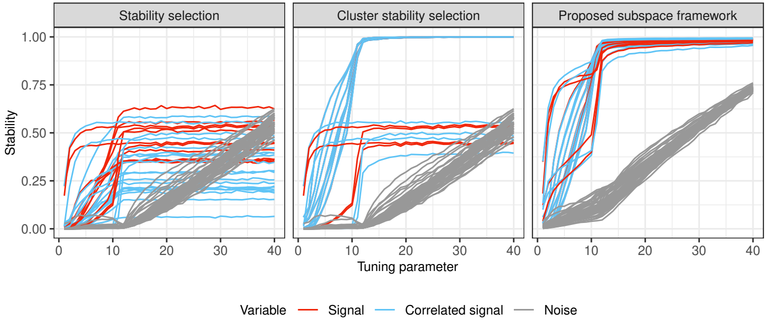
For the given fixed design matrix $X \in \mathbb{R}^{n \times p}$,

- **map** a selection set S onto its **column space** $\mathrm{col}(X_S)$
- $\mathrm{col}(X_S)$ is a subspace living in $\mathbb{R}^n$
- $\mathcal{P}_{X_S}(y)$ is the best linear prediction of y by S

Our contribution: a subspace perspective

For the given fixed design matrix $X \in \mathbb{R}^{n \times p}$,

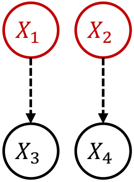
- **map** a selection set S onto its **column space** $\mathrm{col}(X_S)$
- $\mathrm{col}(X_S)$ is a subspace living in $\mathbb{R}^n$
- $\mathcal{P}_{X_S}(y)$ is the best linear prediction of y by S



Why Are Subspaces Useful?

In the toy example:

- Sets $\{1\}$ and $\{3\}$ share no features, yet $\mathrm{col}(X_1) \approx \mathrm{col}(X_3)$
- Sets $\{1, 2\}$ and $\{1, 4\}$ overlap in only one feature, yet $\mathrm{col}(X_{1,2}) \approx \mathrm{col}(X_{1,4})$

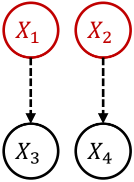


Example: $X_3 \approx X_1$, $X_4 \approx X_2$

Why Are Subspaces Useful?

In the toy example:

- Sets $\{1\}$ and $\{3\}$ share no features, yet $\mathrm{col}(X_1) \approx \mathrm{col}(X_3)$
- Sets $\{1, 2\}$ and $\{1, 4\}$ overlap in only one feature, yet $\mathrm{col}(X_{1,2}) \approx \mathrm{col}(X_{1,4})$



Example: $X_3 \approx X_1$, $X_4 \approx X_2$

Key Idea

Subspace alignment captures similarity even when sets differ

Why Are Subspaces Useful?

In the toy example:

- Sets $\{1\}$ and $\{3\}$ share no features, yet $\mathrm{col}(X_1) \approx \mathrm{col}(X_3)$
- Sets $\{1,2\}$ and $\{1,4\}$ overlap in only one feature, yet $\mathrm{col}(X_{1,2}) \approx \mathrm{col}(X_{1,4})$

$$\begin{array}{cc} X_1 & X_2 \\ \downarrow & \downarrow \\ X_3 & X_4 \end{array}$$

Example: $X_3 \approx X_1$, $X_4 \approx X_2$

Key Idea

Subspace alignment captures similarity
even when sets differ

Implications:

- Higher $\mathrm{corr}(X_1, X_3) \Rightarrow$ stronger alignment: $\mathrm{col}(X_1) \approx \mathrm{col}(X_3)$
- Subspace alignment varies **smoothly** with correlation
- If features are **orthogonal**, then alignment $=$ exact set overlap

Subspace True positives and False Positives: Definitions

Definition (true positive, false positive)

Let $S^\star$ be the true set of features, and $\widehat{S} \subseteq \{1, \ldots, p\}$ be the estimated set. Then:

$$\mathrm{TP}(\widehat{S}, S^\star) := \mathrm{trace}\left(\mathcal{P}_{\mathrm{col}(X_{\widehat{S}})}\mathcal{P}_{\mathrm{col}(X_{S^\star})}\right),$$

$$\mathrm{FPE}(\widehat{S}, S^\star) := |\widehat{S}| - \mathrm{TP}(\widehat{S}, S^\star)$$

Subspace True positives and False Positives: Definitions

Definition (true positive, false positive)

Let $S^\star$ be the true set of features, and $\widehat{S} \subseteq \{1, \ldots, p\}$ be the estimated set. Then:

$$\mathrm{TP}(\widehat{S}, S^\star) := \mathrm{trace}\left(\mathcal{P}_{\mathrm{col}(X_{\widehat{S}})} \mathcal{P}_{\mathrm{col}(X_{S^\star})}\right),$$
$$\mathrm{FPE}(\widehat{S}, S^\star) := |\widehat{S}| - \mathrm{TP}(\widehat{S}, S^\star)$$

- TP measures how well the selected subspace aligns with the true subspace
- FPE counts the dimensions in $\widehat{S}$ that don't align with the true subspace
- Always: $\mathrm{TP} \leq \min\{|\widehat{S}|, |S^\star|\}$

Subspace True positives and False Positives: Definitions

Definition (true positive, false positive)

Let $S^\star$ be the true set of features, and $\widehat{S} \subseteq \{1, \ldots, p\}$ be the estimated set. Then:

$$\mathrm{TP}(\widehat{S}, S^\star) := \mathrm{trace}\left(\mathcal{P}_{\mathrm{col}(X_{\widehat{S}})}\mathcal{P}_{\mathrm{col}(X_{S^\star})}\right),$$
$$\mathrm{FPE}(\widehat{S}, S^\star) := |\widehat{S}| - \mathrm{TP}(\widehat{S}, S^\star)$$

- TP measures how well the selected subspace aligns with the true subspace
- FPE counts the dimensions in $\widehat{S}$ that don't align with the true subspace
- Always: $\mathrm{TP} \leq \min\{|\widehat{S}|, |S^\star|\}$

Geometric interpretation

$$\mathrm{TP} = \sum_i (\cos\theta_i)^2 \quad \mathrm{FPE} = |\widehat{S}| - \sum_i (\cos\theta_i)^2$$

where θ_i are the **principal angles** between $\mathrm{col}(X_{\widehat{S}})$ and $\mathrm{col}(X_{S^\star})$.

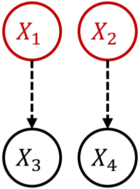
Subspace True Positives and False Positives: Example

$$\mathrm{TP} = \sum_i (\cos\theta_i)^2 \quad \mathrm{FPE} = |\widehat{S}| - \sum_i (\cos\theta_i)^2$$

where θ_i are the **principal angles** between $\mathrm{col}(X_{\widehat{S}})$ and $\mathrm{col}(X_{S^\star})$.

- Let $S^\star = \{1, 2\}$ and we estimate $\widehat{S} = \{3, 4\}$
- In this example:

$$\theta_1 \approx 0^\circ, \quad \theta_2 \approx 0^\circ \quad \Rightarrow \quad \mathrm{TP} \approx 2, \quad \mathrm{FPE} \approx 0$$



Example: $X_3 \approx X_1$, $X_4 \approx X_2$

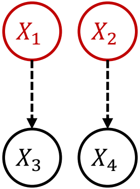
Subspace True Positives and False Positives: Example

$$\mathrm{TP} = \sum_i (\cos\theta_i)^2 \quad \mathrm{FPE} = |\widehat{S}| - \sum_i (\cos\theta_i)^2$$

where θ_i are the **principal angles** between $\mathrm{col}(X_{\widehat{S}})$ and $\mathrm{col}(X_{S^\star})$.

- Let $S^\star = \{1, 2\}$ and we estimate $\widehat{S} = \{3, 4\}$
- In this example:

$$\theta_1 \approx 0^\circ, \quad \theta_2 \approx 0^\circ \quad \Rightarrow \quad \mathrm{TP} \approx 2, \quad \mathrm{FPE} \approx 0$$



Example: $X_3 \approx X_1$, $X_4 \approx X_2$

More generally:

- principal angles vary smoothly
- if subspaces are orthogonal:

$$\mathrm{TP} = |\widehat{S} \cap S^\star|, \quad \mathrm{FPE} = |\widehat{S} \setminus S^\star|$$

Reduces to classical notions

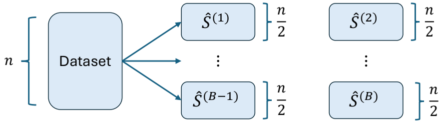
Subspace Stability (2nd challenge): Setup

Goal

Measure how consistently subspaces are selected across random subsamples.

Apply Lasso or ℓ_0 regression to:

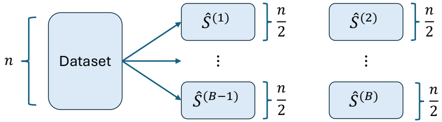
- B subsamples of size $\lfloor n/2 \rfloor$
- Each produces a set $\widehat{S}^{(\ell)}$



Subspace Stability (2nd challenge): Setup

Apply Lasso or ℓ_0 regression to:

- B subsamples of size $\lfloor n/2 \rfloor$
- Each produces a set $\widehat{S}^{(\ell)}$



Goal

Measure how consistently subspaces are selected across random subsamples.

Definition (stability)

Given B subsamples:

$$\mathcal{P}_{\text{avg}} := \frac{1}{B} \sum_{\ell=1}^{B} \mathcal{P}_{\text{col}(X_{\widehat{S}(\ell)})}.$$

Then for any set S,

$$\begin{aligned} \pi(S) &:= \sigma_{|S|}(\mathcal{P}_{\text{col}(X_S)} \mathcal{P}_{\text{avg}} \mathcal{P}_{\text{col}(X_S)}) \\ &= \min_{\substack{z \in \text{col}(X_s) \\ \|z\|_2 = 1}} \frac{1}{B} \sum_{\ell=1}^{B} \|\mathcal{P}_{\text{col}(X_{\widehat{S}(\ell)})}(z)\|_2^2 \end{aligned}$$

What Does Stability $\pi(S)$ Measure?

Stability of set S with respect to estimates $\{\widehat{S}^{(\ell)}\}_{\ell=1}^{B}$

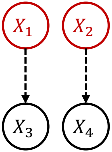
$$\pi(S) := \sigma_{|S|}(\mathcal{P}_{\mathrm{col}(X_S)}\mathcal{P}_{\mathrm{avg}}\mathcal{P}_{\mathrm{col}(X_S)}).$$

with $\mathcal{P}_{\mathrm{avg}} := \frac{1}{B}\sum_{\ell=1}^{B}\mathcal{P}_{\mathrm{col}(X_{\widehat{S}(\ell)})}$

- $\pi(S) \in [0, 1]$
- $\pi(S)$ captures the **lowest alignment** of any direction in $\mathrm{col}(X_S)$ with $\mathcal{P}_{\mathrm{avg}}$
- $\pi(S)$ is the smallest squared cosine between any direction in $\mathrm{col}(X_S)$ and the average of the subsample subspaces.
- High $\pi(S)$ means that $\mathrm{col}(X_S)$ is reliably recovered across subsamples

Back to Toy Example: $\pi(\{1\}) \approx \pi(\{3\}) \approx 1$

- $\sim$50% of subsamples select X_1
- $\sim$50% select X_3
- But $X_3 = X_1 + \delta$ — nearly the same direction
- So all subsamples nearly capture the X_1 and X_3 direction!
- Therefore: $\boxed{\pi(\{1\}) \approx \pi(\{3\}) \approx 1}$

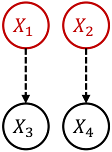


$X_3 = X_1 + \delta$ (δ small)

Both signal features and highly correlated non-signal features have high stability π!

Back to Toy Example: $\pi(\{1,3\}) \approx 0$

- $\sim$50% of subsamples select X_1
- $\sim$50% select X_3
- $\mathrm{col}(X_{1,3})$ includes directions X_1 and δ
- But no subsample selects both X_1 and X_3 together
- So $\mathcal{P}_{\mathrm{avg}}$ misses the δ direction
- Therefore: $\boxed{\pi(\{1,3\}) \approx 0}$



$X_3 = X_1 + \delta$ (δ small)

Redundant features in S lead to small stability $\pi(S)$

Table of Contents

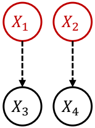
FSSS algorithm (3rd challenge)

We propose an algorithm "Features Subspace Stability Selection" (FSSS)

FSSS algorithm

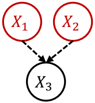
- Input: Matrix $\mathcal{P}_{\mathrm{avg}}$, and stability threshold $\alpha \in (1/2, 1)$
- Sequentially add variables:
  - Start from the null model $\varnothing$
  - Add X_j to the current model S if $\pi(S \cup \{j\}) \geq \alpha$
  - Stop until no new features can be added
- Output: one selection set $\widehat{S}$

Toy example ♠



- Any output $\widehat{S}$ is at least α-**stable**: $\pi(\widehat{S}) \geq \alpha$.
- Each run may given different selection sets:
  - For example ♠, can return $\{1,2\}$, $\{1,4\}$, $\{2,3\}$, $\{3,4\}$
  - For example ♣, can return $\{1,3\}$, $\{1,2\}$, $\{2,3\}$

Toy example ♣



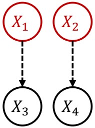
FSSS algorithm (3rd challenge)

We propose an algorithm "Features Subspace Stability Selection" (FSSS)

FSSS algorithm

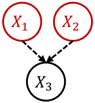
- Input: Matrix $\mathcal{P}_{\mathrm{avg}}$, and stability threshold $\alpha \in (1/2, 1)$
- Sequentially add variables:
  - Start from the null model $\varnothing$
  - Add X_j to the current model S if $\pi(S \cup \{j\}) \geq \alpha$
  - Stop until no new features can be added
- Output: one selection set $\widehat{S}$

Toy example ♠



- Any output $\widehat{S}$ is at least α-**stable**: $\pi(\widehat{S}) \geq \alpha$.
- No redundant features can be included:
  - For example ♠ (as well as example ♣), $\{1, 2, 3\}$ can not be returned since $\pi(1, 2, 3) \approx 0$!

Toy example ♣

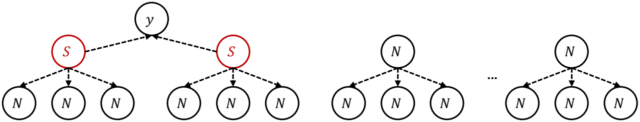


FSSS algorithm

Now that the models returned by FSSS are stable...
How accurate (FPE) can they be?

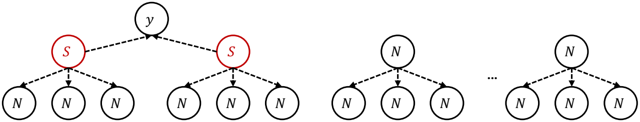
FSSS algorithm

Now that the models returned by FSSS are stable...
How accurate (FPE) can they be?



FSSS algorithm

Now that the models returned by FSSS are stable...
How accurate (FPE) can they be?



Theorem (False positive error control)

Under the clustering setup (as above), for any $\widehat{S}$ returned by FSSS with $|\widehat{S}| \leq s_0$,

$$\mathbb{E}\left[\mathrm{FPE}(\widehat{S}, S^\star)\right] \leq \frac{p(\gamma + b_n)^2}{2\alpha - 1} + a_n.$$

- The term γ is a **quality term** of the base procedure, with $\gamma \approx s_0/p$.
- a_n and b_n are **slackness** terms from dealing with perturbation and decomposing projection matrix.

FSSS algorithm

Now that the models returned by FSSS are stable...
How accurate (FPE) can they be?

Theorem (False positive error control)

Under the clustering setup (as above), for any $\widehat{S}$ returned by FSSS with $|\widehat{S}| \leq s_0$,

$$\mathbb{E}\left[\mathrm{FPE}(\widehat{S}, S^\star)\right] \leq \frac{p(\gamma + b_n)^2}{2\alpha - 1} + a_n.$$

- The term γ is a **quality term** of the base procedure, with $\gamma \approx s_0/p$.
- a_n and b_n are **slackness** terms from dealing with perturbation and decomposing projection matrix.

- Stability-selection-type upper bound
- orthogonal features $\Rightarrow$ terms a_n and b_n **vanish** $\Rightarrow$ UBD becomes $\frac{s_0^2}{p(1-2\alpha)}$

Experiments

A synthetic dataset including
...multiple **cluster** blocks, **parent-children** blocks, and **independent** features.

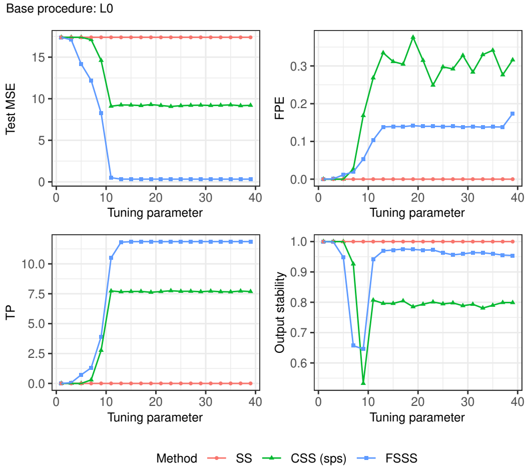
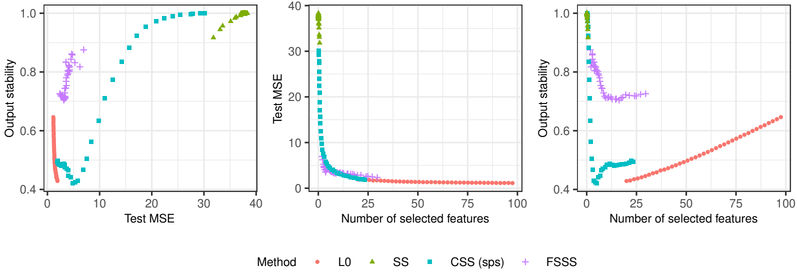


Table of Contents

Gene expression in breast cancer

A gene expression dataset with $n = 189$, and $p = 1111$

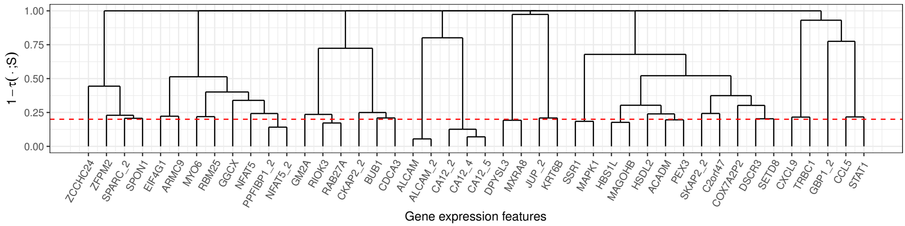
- Step 1: Performance comparison



- Step 2: Repeat FSSS algorithm to obtain 45 stable models

Gene expression in breast cancer

Among the 45 stable models:



Even more:
suppose that a domain expert give me some interesting features sets

We can do:

- rank those sets, or features, based on stability
- calibrate the commonality among those sets
- calibrate the substitutability among those sets

Summary

For model selection among highly-correlated predictors:

- Proposed a subspace framework:
  - Re-defined **true positive** & **false positive error**
  - Re-defined **stability**
  - Designed an algorithm **FSSS** that outputs stable models
- Future work:
  - More efficient FSSS algorithm
  - Extension to the non-linear setup: generalized additive models

Thanks for listening!

Check out the **full article**:
https://arxiv.org/abs/2505.06760

Check out the **package**:
https://github.com/Xiaozhu-Zhang1998/substab